

Functional Requirements for Implementing the GeoArchives

A DRAFT #7

Jim Henderson, Maine State Archivist

David Weaver Commentary in Bold Blue.

OVERVIEW

The creation of the GeoArchives implies that several functional requirements be met. An enumeration of these requirements will help guide the project and provide a basis to assess progress. As noted in the project proposal, the goal is to:

create a digital system, based on current archival research and recommended standards, . . . for maintaining State of Maine GIS records having permanent value. . . . The GeoArchives will insure that State or local government archival records held by the GeoLibrary, as well as other archival GIS records held by State agencies but not integrated into the GeoLibrary, will be retained permanently and will be accessible to researchers through the GeoLibrary.

A list of "functional requirements" for implementing the GeoArchives appears below. The Reference Model for an Open Archival Information System (OAIS) may provide guidance on how to achieve our goals.[1] However, the OAIS appears oriented to simpler data objects than those in GIS systems. The "Dublin Core" of metadata elements may also offer useful insights.[2] (See attached summary.) The term "data object" used here, following the OAIS model, means layers, features, attributes, image files, etc. "Records" will refer to paper or digital textual or graphic material that document the operation of the GIS "system" or document layers appraised as archival.[3]

FUNCTIONAL REQUIREMENTS

1. Establish the appraisal level
2. Determine the administrative records to be preserved
3. Determine the types of data objects to be preserved
4. Appraise each set of records and objects for their archival value
5. Determine what additional metadata is needed
6. Determine in what formats data objects will be retained
7. Establish provisions to insure the integrity and security of data objects
8. Determine how archival objects can be retrieved to produce gis products
9. Determine state archives administrative arrangements needed

STATUS OF FUNCTIONAL REQUIREMENTS

1. THE APPRAISAL LEVEL

Argument

Several options might be considered for preservation: all State GIS data objects and GIS systems; selected GIS administrative records; products of GIS systems, such as maps produced from selected layers to aid in government decision-making; all layers and image sets; selected image sets and selected layers including all feature attributes; selected layers and selected feature attributes.

Archival GIS data objects should be retained in a form that is likely to survive changes in GIS systems, thus the systems themselves need not be retained.

- **Some software, hardware and applications should be retained for historical reasons. Other software and applications should be preserved because it will allow future users to view the data in a similar way that the original users did.**

However, policies and standards governing creation of GIS data will inform future users of the reliability and context of those data objects. Not all GIS data objects are likely to be permanently valuable. Not all features within a layer are likely to be permanently valuable.

- **I do not fully agree with this. Data that meets the criteria of the GeoLibrary generally should qualify for inclusion in the GeoArchives both because it was the best GIS data available for a period of time (Informational), and because much of it would have been used by state agencies as part of a decision-making process (Evidential). In addition it will be expensive and time consuming to selectively delete specific features from a data set (or series).**

Boundaries of this Project.

Products such as maps, created to aid in government decision-making, should be scheduled at the creating agency level. Scheduling of these products is not a task for this project.

- **GeoArchives staff, or other Maine archives staff, may be called to provide advice on the scheduling and archiving of potential items and series of GIS related products.**

Given potential complexities arising from non-State GIS data created by municipalities, counties, or other entities, this project will seek solutions for State agency GIS records 1) in the GeoLibrary or, 2) at our discretion, candidates for inclusion in the GeoLibrary.

- **This is very important, as non-state data generated at the local level is unlikely to be permanently archived unless it physically resides on GeoLibrary servers or is subsequently collected by the GeoArchives.
These data will need assistance to bring metadata, formats, etc to at least minimum standards for archival access and utility**

Conclusion

For administrative records, those critical for informed future use of GIS data objects preserved through this project will be appraised as permanently valuable.

For GIS data objects, appraisal will be at the layer and image set levels. Images will normally be appraised as complete sets.

Two layers have been selected for implementation in this project: 8/24/04

1. E-911 Roads Abstract:

E911RDS contain updated road centerline and road name data for Maine at 1:24,000 scale. Data is statewide and divided by minor civil divisions. The data set was developed at the Maine Office of GIS by the Enhanced 911 project (E911) from USGS 1:24,000 digital roads data and is in Arc/Info layer format. E911RDS data was developed to implement the Enhanced 911 project in Maine. This data contains up-to-date road names and address ranges for the State of Maine. Data at this scale would be useful for planning, utility, development and various other applications. Not for use for scales greater than 1:24,000.

This is a

2. Maine Township Boundaries at 1:24,000 scale Abstract:

METWP24 depicts political boundaries, common town names, and geocodes for Maine at 1:24,000 scale. The layer was created from USGS, 7.5 minute map series, town boundaries. METWP24 was created to show political boundaries, common town names

and Maine geocodes at 1:24,000 scale. Data at this scale is suitable for detailed studies and local planning. Not for use in scales greater than 1:24000.

- **These data sets can be considered good “pilot” data to test, prove, and improve the plan for archiving GIS data sets. This pilot should use data sets that are considered by the Maine GeoLibrary to be operational, to have been fully reviewed and accepted, and with complete metadata.**

Additional pilot data could be:

- **a image data set (out-of-date orthophoto or scanned topographic sheet**
- **an out-of-date vector data sets such as land use**
- **a data set of point features**
- **a sensitive data set, such as a cancer registry where a certain attribute needs to be taken out of features in the data set before making it available to researchers and/or the public (see #3 below)**
- **The pilot project should be structured to address both preservation and access issues as well as metadata. To do this properly, a component of the pilot needs to develop a prototype access system to be tested in actual use. To be truly useful, the user needs to be able to not only easily find the data, but to use it within a robust GIS user interface, which includes statistical and visual combination with current and other GeoArchived data; visualization; clear (and perhaps user defined) symbology; and statistical and graphic output.**
- **Carefully monitoring this pilot will also help the Archives to realistically gauge the complexity and cost of populating the GeoArchives with new data sets. This will be invaluable for future planning and implementation.**

Remaining Tasks

- Consider the addition of other layers to test the adequacy of the prototype with a range of layer types. Nominees include Land Use Regulation Commission & Revenue Service land parcel layers;

Because of its complexity, it is likely that archiving parcel data that is continually updated will be one of the most challenging data sets to manage in the GeoArchives. This is especially the case if transactional data is captured in an SDE-type format. This data should also have backup ‘snapshots’ made on a periodic basis.

- Identify image sets for appraisal: film-based photography, satellite imagery, State copies of USGS ortho-imagery.

- **There is good potential for cooperative agreements with the USGS and other federal agencies to take responsibility to archive and share data sets via internet protocols, e.g. web services. These cooperative agreements, if created, will be a win-win-win for both agencies and the user community, as redundant data sets would not have to be archived and managed.**
- **However, having data in only one archive IS a risk for the long-term survivability of any data set. Doing this presupposes that the archival custodian has a completely foolproof system for preserving, migrating and making the data accessible.**

2. GIS ADMINISTRATIVE RECORDS TO BE PRESERVED

Argument

Documenting the context in which GIS objects are created and maintained is critical to future users, both to insure functionality and to allow intelligent interpretation. Records documenting the GeoLibrary and its standards for donors will provide an overview of how GIS records have been, and will be, administered in Maine State government. Other records in the creating agencies will document sources of data, programs for which the objects were created, etc.

Conclusion

The Archives' Records Management Analyst should insure these administrative records are scheduled. The records will include those documenting operating rules and standards, administrative relationships, metadata requirements for agencies, etc.

Remaining Tasks

The Records Management Analyst should meet with appropriate GeoArchives team members, and with creating agency staff responsible for layers or images appraised as archival, to insure these records are scheduled. The Analyst should, for the GIS system and for each layer or image appraised as archival, report to the GeoArchives Team which records are already scheduled as archival, whether they are already in the Archives, and which records need to be scheduled.

- **I heartily agree.**
- **For the record, the universe of types of information and objects to be retained SOMEWHERE under a Maine GeoArchives program could/should include:**
 - **Digital GIS data sets as discussed in the NHPRC proposal- to be readable in the future (probably starting out in their GIS native format as currently defined/proposed by MeGIS)**
 - **Query and access software eventually including 'virtual machines' running on future computers-this will show best how digital data was used at the time of first use including standard cartographic symbology. This will help future researchers see the data as was seen by state scientists, planners and decision makers**
 - **Sample (standard) queries as set up for various departments and programs**
 - **Statewide data sets combined into a hardcopy atlas**
 - **Sample standard maps (not a whole atlas) as created by various departments**
 - **Key maps (hardcopy and digital views) used for critical decision-making (this will help future researchers understand the maps used when the decisions were being made)**
 - **Various hardware likely to amuse future students of GIS history (e.g. EGA monitors, digitizing tables)**
 - **Evidential records of the Office of GIS; other GIS operations at all levels of government and selected other entities**

3. DETERMINE THE TYPES OF DATA OBJECTS TO BE PRESERVED

Argument

Given the conclusion in #1 above, objects essential to the preservation of layers and images must be preserved. Preservation of all elements of a layer may be more efficient than parsing its separate elements.

Conclusion

Based on the hypothesis that preservation of all elements of a layer may be more efficient, all features and attributes within a layer will be retained, even if some might not be considered permanently valuable in

isolation. Image sets will be retained in a format appropriate compatible with use in the GIS environment.

- **I strongly agree with this approach- keep data sets whole and intact. There may need to be consideration for public safety and privacy issues. However, it is highly probable that any data sets that have been included in the GeoLibrary will already be filtered and/or altered to protect these two important issues.**

In some cases it may be advisable to maintain all features in a layer, but make a change, e.g.:

- **generalize the location of critical resources or individuals so as not to fully disclose that information, yet allow the data to be used for statistical aggregations**
- **delete one or more attributes in a data set (e.g. individual name or specific medical condition)**

Remaining Tasks

?

- **Go through the process of stripping private, sensitive attributes from a data set as part of a pilot phase data set conversion to the GeoArchives.**
- **Research and discuss likely 'fringe' data sets, e.g. those not firmly under the wing of the GeoLibrary, but deemed important for various reasons**
- **Data that be disguised, generalized or deleted for public safety/homeland security reasons. For example**

4. APPRAISE EACH SET OF RECORDS AND OBJECTS FOR THEIR ARCHIVAL VALUE

Argument

The following principles and procedures guide the Maine State Archives in appraising records and objects for their archival value. Records having administrative, legal or fiscal value are retained for as long needed by the creating agency or other agencies.

Administrative: necessary to perform mandated functions.

Legal: necessary to insure that legal mandates were met or that legal rights of organizations or individuals are protected.

Fiscal: necessary to assure others (Audit, Budget Office, Controller, and Legislature) that funds have been properly expended. While often these "primary" values are time limited, those that last indefinitely require permanent (archival) retention. When primary values have ceased for official records, they are appraised for their evidential or informational value to others.

Evidential: the record of mission, organization, and functioning; the significant facts in an agency's existence - its purpose, patterns of action, policies, procedures, and achievements.

Informational: the information contained in the records, regardless of the creating agency; for example, information about environmental status (water or air quality) or health conditions (cancer registry). Determining informational value is based on several factors and is ultimately a judgment to be made by the State Archivist and the Archives Advisory Board.

Uniqueness: Is the information not found elsewhere; elsewhere but incomplete; elsewhere but dispersed; do the records form an important link to explaining other archival records?

Format: Are the records accessible and usable, or can they reasonably be made accessible and usable?

Content: Is there a substantial amount of well-documented information in the records?

Audience: Is there a body of known or potential users, e.g., policy makers, environmental organizations, academic disciplines, private industry and commercial groups, consumer groups, public service client groups, genealogists?

Each data layer and image set will be appraised for its archival value.

Conclusion

The principles outlined above will be considered in appraisal decisions.

Remaining Tasks

Specifically related to the two selected layers, E-911 and METWP24, the Archives will develop a model assessment justifying their selection as archival.

- **Geographic data, more than most data, may be Evidential information to the custodial/creation agency, but important Informational data to a number of other agencies and other GeoArchives users. Appraisal must keep this wider picture of data use in mind during appraisal.**
- **Many of the data sets that will be appraised for the GeoArchives will be mainly Informational in nature.**

5. DETERMINE WHAT ADDITIONAL METADATA IS NEEDED

Argument

Not all GIS metadata is likely to serve archival needs.

Conclusion

The existing metadata must be compared to required archival metadata to determine additional metadata needs.

Remaining Tasks

Determine what additional metadata is needed for preservation, retrieval, description, provenance, fixity (data integrity), and security. How can the archival metadata suggested by the Dublin Core be accommodated?

Suggested Additional Metadata Fields:

- **A record of the data set's life as an active data set within the Maine GeoLibrary**
- **Information linking the data set to previous and subsequent incarnations of the similar data within the GeoLibrary**
- **Possible editorial/oral history of the data set and its usage by staff of creating agency**

6. DETERMINE IN WHAT FORMATS DATA OBJECTS WILL BE RETAINED

Argument

Select the most "open" formats and standards. Since digital records by their nature are dependent on digital systems for creation, storage, and retrieval; and since all current systems rely on proprietary software from operating systems to programs for creation and retrieval, permanent retention of these records requires an approach that offers a high probability of migration to successor systems. Retention strategies may differ between layers that are updated at defined intervals (e.g., annually or less frequently) and those that are continuously updated (e.g., roads).

Since the data for GIS layers is retained using relational database systems in platform independent "tables," the data and related metadata may be exported to other contemporary and successor relational database systems. Imagery is received from several sources, but largely from the United State Geological Survey (USGS). Photographic film images are created and retained by USGS, which then creates ortho-rectified

digital images from contiguous sets of film images. USGS then integrates these ortho-rectified digital images into a single mosaic of ortho-imagery, which is acquired and owned by the State of Maine.

Conclusion

Retain data in relational database structures. In the case of State of Maine GIS layers, current data is retained in Oracle® databases. Project GIS staff confirms that GIS layers in standard table-based Oracle® databases using ESRI's SDE® product are stored in a manner that the data may be retrieved for use without SDE®.

Retain layers that are updated at defined intervals (e.g., annually or less frequently) as "snapshots" immediately before the beginning of the next update interval. Develop a system to document changes to layers that are continuously or frequently updated by pursuing promising proprietary technologies, or if they do not appear promising, developing an open alternative. In any event, the initial data along with the documented changed data must reside in a standard table-based relational database.

Retain images in the current digital ortho-imagery mosaic format used for GIS storage.

Determine whether the film images and the ortho-rectified digital images that constitute the statewide mosaic are retained permanently by USGS. If so, they need not be scheduled as archival in Maine. If not, the project should assess whether future migration would be compromised by not having access to these images.

Remaining Tasks

Establish clear rules or guidelines for layer metadata and retention structures.

Participate as a beta test site for ESRI's SDE 9.1® product for archiving frequently updated layers, as recommended in the "Technical Recommendations" dated November 15, 2004.

- **This section is clear and well thought out.**

7. ESTABLISH PROVISIONS TO INSURE THE INTEGRITY AND SECURITY OF DATA OBJECTS

Argument

Archival records must be secure from both destruction and from unauthorized changes in content (integrity). The latter insures their authenticity as reliable public records. This requires developing clear rules for transferring layers to Maine State Archives jurisdiction and "ownership."

Once in "archival status" no undocumented changes will be permitted. This is to provide current and future users with a resource to reconcile discrepancies among versions of the same dated layer.

Proposal: For layers retained as Archival "snapshots" at defined intervals, the following rules might apply:

1. Changes made before the "snapshot date" and before "publication" for general use (through the GeoLibrary or otherwise) are at the discretion of the creating agency;
2. Changes made before the "snapshot date" and after "publication" for general use (through the GeoLibrary or otherwise) are at the discretion of the creating agency, but must be documented to include the status of the changed data before the change.
3. Changes made in the Archival "snapshot" after the "snapshot date" are at the discretion of the Maine State Archives, in consultation with the creating agency, and must be documented to include the status of the changed data before the change.
4. Layers that enter archival status will be named to indicate their snapshot date, such as metwp24-050630 for the June 30, 2005 snapshot of metwp24.

5. Metadata in each layer will indicate the reason for selecting the snapshot date. E.g., "annual, end of fiscal year"; "annual, end of calendar year": "18-month cycle beginning January 1, 2005"; "special event, substantial mandated realignment of boundaries effective May 15, 2005."

- **Excellent**

Proposal: For imagery scheduled as archival, the following rules might apply:

1. All documentation related to the creation of the imagery must be appraised for potential archival retention.
2. A system for identifying and locating images that cover a specific location must be documented or created.

- **This will only be an issue with non-georeferenced data and images (e.g. raw aerial photography both vertical and oblique). Georeferenced images will be very easily locatable via standard GIS software queries.**

3. To insure integrity of a GIS imagery database, the database should either be "locked" upon receipt by the GeoLibrary or a copy made for the GeoArchives. If a replacement image is inserted into the database, that replacement must be documented to include the status of the changed data before the change.

- **Imagery data will be read-only as a matter of standard data management practices, both in the GeoLibrary and the GeoArchives. This is also true of feature based data sets, e.g. point, line and polygon features.**

Conclusion

Create a "snapshot" for all layers appraised as archival and updated annually or less frequently, as recommended in the "Technical Recommendations" dated November 15, 2004.

Appraisal of these layers must include:

- 1) assessing the creating agency's current update schedule,
- 2) insuring that the schedule is documented in its metadata, and
- 3) scheduling the update interval by the Archives Advisory Board.

Remaining Tasks

Provisions must be developed to either insure that archival objects 1) are not changed, or 2) if changed, are done so with proper authorization and complete documentation. Determine the technical requirements for insuring security and integrity.

Develop clear rules for 1) transfer to archival status, 2) determining how and in what circumstances "corrections" may be made, 3) determine how to document corrections made while in archival status.

- **Don't destroy the ability to see actual data that was used during an era (in the Evidential sense).**

Confirm with the Archives that the levels of integrity maintenance and security are acceptable.

Review, revise, and adopt rules suggested in the Argument above.

Implement these requirements in the prototype.

8. DETERMINE HOW ARCHIVAL OBJECTS CAN BE RETRIEVED TO PRODUCE GIS PRODUCTS

Argument

Archival preservation exists to provide public access now and in the future.

Access should be relatively convenient for GIS non-experts.

Since electronic digital records by their nature are dependent on electronic digital systems for retrieval; and since current systems rely heavily on proprietary software, access to these records requires an approach that insures delivery of GIS data and imagery that is accessible to potential users.

Proprietary, or possibly open-source, software may be used to query GIS data from a standard tables-based relational database and to prepare it for convenient access for potential users through the GeoLibrary system. Users should be able to access this data without having a particular brand of proprietary software.

This project should not, unless resources allow, develop value-added convenient access products for users. The primary focus should be on strategies for permanent retention. Thus providing Internet Mapping System access is not a priority unless it offers a cost-effective (to the project) alternative to other solutions.

- **I disagree with this, access is a vital goal of this project. Working closely with the GeoLibrary and staff should make robust access affordable without ‘reinventing the wheel’.**
- **Based on current trends, it is highly likely that Internet based access solutions (or their direct successors) will be the most cost effective and useful ways to access the GeoArchived data sets.**

Conclusion

We should determine how archival objects could be retrieved and integrated with current objects and other archival objects to produce GIS products, such as maps displaying several layers. A clear understanding about this should be developed along with the preservation model.

- **This is a vital concept. Access to GeoArchived data MUST include the ability to view and analyze the data in the context of current and other GeoArchived data.**

We should develop access tools that do not require the user to obtain proprietary software. One model would be to extract GIS data from a relational database and serve it to the Internet relying on some proprietary software, then allowing access by remote users through an open-source non-proprietary interface that would permit use of the data by a variety of available GIS tools.

Remaining Tasks

Develop the data extraction and access tools.

Determine if XML provides advantages for transfer of GIS data to users.

- **Metadata may be XML format, which is an emerging standard in GIS, libraries and archives.**

9. DETERMINE STATE ARCHIVES ADMINISTRATIVE ARRANGEMENTS NEEDED TO INSURE THE SURVIVAL AND USE OF ARCHIVAL OBJECTS

Argument

The Maine State Archives, in association with the GeoLibrary, needs to develop administrative procedures to maintain intellectual and physical control over GIS objects that are appraised as archival. We may wish to measure the success of the GeoArchives by determining if it meets

Functional Requirements For Recordkeeping Systems, as suggested by David Bearman and others:

Compliant Insure that the system and rules created to maintain the GeoArchives operate as designed.

Responsible Insure the GeoArchives has policies, assigned responsibilities, and formal methods for effective management.

Credible The GeoLibrary and Geo Archives must monitor the quality of information being input and insure that information is accurate, documented and consistent with policies.

Complete: Records incorporate or link to information about the context of their creation, e.g., the relevant administrative records.

Authentic: The system must validate records creators and/or authorizers to insure information is authentic.

Sound: Record integrity is protected from accidental or purposive damage or destruction and from any modification after they have been placed in archival status.

Auditable: GeoArchives' controls preserve auditability of interactions external to the system (such as during media migration or transfer).

- **Conversion and processing of data sets within the ESRI ArcGIS environment will generate a step-by-step history of these modifications.**

Available: The GeoArchives system must document all logical archival records it contains, indicate the terms under which they are available for research, and retrieve them for authorized users.

- **And these should be available to be retrieved for users within a GIS software environment**

Exportable: Record content, structural representation and representation of context must be exportable, in standard protocols.

- **For appropriate and approved uses, the data must be able to be exported to other GIS, e.g. sent to a state or federal agency as part of a grant application, etc. Physical export of data will be minimized and mitigated by the use of web services, which allow for the use of data within a GIS application over the internet without physically copying and moving the entire data set.**

Renderable: The GeoArchives system must present records to allow views in effect at the time any record was in existence.

Redactable: The GeoArchives system must support delivery of redacted, summarized, or censored copies and keep records of the version released.

- **This means GIS analysis and viewing software**

Conclusion - The project should generate a set of administrative issues that need to be resolved, in addition to the technical issues, and address each issue.

Remaining Tasks

Identify all such issues, including the following:

1. Track all GIS objects appraised as archival and provide access to that information.
2. Insure that "ownership" of these objects is clearly attributed to the Maine State Archives.
3. Log frequency of access to these objects for statistical purposes.

4. Determine what costs should be born by the Archives for storage of these objects.
5. Determine financial support mechanisms for maintaining archival objects.

This is an evolving document that may not contain all considerations necessary for the completion of the GeoArchives Project.

Metadata Guidelines (Based on Dublin Core Schema)

These may be most appropriate for documenting the layer.

Title 245A name given to the resource. Title is a name by which the resource is formally known.

Creator 100, 110, 111, 700, 710, 711, 720 An entity primarily responsible for making the content of the resource. Examples of Creator include a person, an organization, or a service.

Subject 600, 610, 611, 630, 650, 653A topic of the content of the resource. Subject is expressed as keywords, key phrases or classification codes that describe a topic of the resource.

Description 500-599, except 506, 530, 540, 546 An account of the content of the resource. Examples of Description include, but are not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

Publisher 260\$a\$b An entity responsible for making the resource available. Examples of Publisher include a person, an organization, or a service.

Date 260\$c A date of an event in the lifecycle of the resource. Typically, Date is associated with the creation or availability of the resource.

Type 655 The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content.

Format 856\$q The physical or digital manifestation of the resource. Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration.

Identifier 856\$u An unambiguous reference to the resource within a given context.

Source 786\$o\$t A reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or in part.

Language 008/35-37 A language of the intellectual content of the resource.
546

Relation 440, 530, 760- A reference to a related resource. Recommended best

practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.

Coverage^{522, 651} The extent or scope of the content of the resource. Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
513\$b, 033\$a

Rights^{506, 540} Information about rights held in and over the resource. Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. These guidelines are intended to assist with the creation of consistent descriptive metadata for archival resources. The metadata schema is based on the Dublin Core schema and uses 14 elements.

- [1] Preservation Metadata and the OAIS Information Model A Metadata Framework to Support the Preservation of Digital Objects, June 2002. Accessed July 26, 2004.
http://www.oclc.org/research/projects/pmwg/pm_framework.pdf
[2] See <http://dublincore.org/documents/dcmi-terms/>
[3] See document asgis04.doc for definitions associated with this project .